Occasional Paper 158 1958

THE ELUSIVE FORMULA OF BEST FIT: A COMPREHENSIVE NEW MACHINE PROGRAM

L. R. GROSENBAUGH

SOUTHERN FOREST EXPERIMENT STATION
PHILIP A. BRIEGLEB, DIRECTOR
Forest Service, U. S. Department of Agriculture

THE ELUSIVE FORMULA OF BEST FIT: A COMPREHENSIVE NEW MACHINE PROGRAM

L. R. Grosenbaugh
Southern Forest Experiment Station

Do you need to fit formulas (or regressions) to data?

For any set of data, how would you like a look at all the least-squares formulas which predict Y using every possible linear combination of 9 or fewer variables? A new electronic machine program now lets you do this cheaply and easily, even though you have no background in mathematics or machines. In addition, the program output for each of the hundreds of formulas includes a value which is useful in comparing reliability or efficiency. Before now, such a capability has existed in theory only--the cost by previously existing machine programs was far too high even for the most inquisitive expert.

What does this mean to you?

Suppose that in managing forest land or carrying out a forest research project you have collected data similar to those shown in figure 1. Maybe Y represents individual sample-tree volumes or plot growth, while X_1, X_2, etc., represent associated measurements such as d.b.h., height, taper, distribution parameters, or functions such as their squares, reciprocals, or logarithms, or joint functions of several of these. Perhaps Y is site index or available soil moisture and the X's are soil or stand variables. Or Y may be cost or value, and the X's may be items thought to influence cost or value. No matter--your problem is to find a formula that will satisfactorily predict Y when only the X's are known, and to discard any column of X's that does not appreciably improve the prediction.

In the past, people have almost never been willing to spend the time and money involved in calculating and comparing the reliabilities of the hundreds of possible formulas which might be fitted by least squares to observations of a dependent variable and 9 or fewer inde-

Figure 1.— Raw or coded data.

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
9	0	0	0	0	0	0	0	0	0
23	0	1	0	0	1	0	0	0	1
17	0	2	0	0	4	0	0	0	8
5	0	3	0	0	9	0	0	0	27
8	0	4	0	0	16	0	0	0	64
9	1	0	0	1	0	0	0	0	0
8	1	1	1	1	1	1	1	1	1
5	1	2	2	1	4	2	4	4	8
15	1	3	3	1	9	3	9	9	27
2	1	4	4	1	16	4	16	16	64
3	2	0	0	4	0	0	0	0	0
3	2	1	2	4	1	4	2	4	1
7	2	2	4	4	4	8	8	16	8
4	2	3	6	4	9	12	18	36	27
9	2	4	8	4	16	16	32	64	64
7	3	0	0	9	0	0	0	0	0
11	3	1	3	9	1	9	3	9	1
7	3	2	6	9	4	18	12	36	8
4	3	3	9	9	9	27	27	81	27
18	3	4	12	9	16	36	48	144	64
9	4	0	0	16	0	0	0	0	0
5	4	1	4	16	1	16	4	16	1
11	4	2	8	16	4	32	16	64	8
14	4	3	12	16	9	48	36	144	27
23	4	4	16	16	16	64	64	256	64
7	5	0	0	25	0	0	0	0	0
18	5	1	5	25	1	25	5	25	1
13	5	2	10	25	4	50	20	100	8
19	5	3	15	25	9	75	45	225	27
29	5	4	20	25	16	100	80	400	64
Totals 322	75	60	150	275	180	550	450	1650	600
Number of sets of observations 30									

Figure 2

SOUTHERN FOREST EXPERIMENT STATION 704 REGRESSION PROGRAM OUTPUT LRG 11- 5-57

IDENTITY NUMBER ITEM	SUM OF SQUARES	0+5	1+6	2+7	3+8	4+9
MEANS		2 10733333	1 25000000	1 20000000	1 50000000	1 91666666
NUMBER PER MEAN		30	30	30	30	30
MEANS		1 60000000	2 18333333	2 15000000	2 55000000	2 20000000
NUMBER PER MEAN		30	30	30	30	30
MAXIMUM MATRIX, SP AND SS						
0	TOTAL =	4 13498667	3 10500000	2 83000000	3 64700000	3 75633335
		3 34900000	4 34096667	4 25730000	5 12909000	4 14510000
1			2 87500000	00000000	3 17500000	3 43750000
		00000000	3 87500000	3 52500000	4 26250000	00000000
2				2 60000000	3 15000000	00000000
		3 24000000	3 55000000	3 60000000	4 22000000	3 92400000
3					3 90000000	3 87500000
		3 60000000	4 40000000	4 32500000	5 14250000	4 23100000
4						4 23741667
		00000000	4 47483334	4 26250000	5 14245000	00000000
5						
		4 10440000	4 22000000	4 26100000	4 95700000	4 42000000
6						
			5 19286667	5 14250000	5 67650000	4 84700000
7						
				5 12720000	5 54900000	5 10500000
8						
					6 25581600	5 38500000
9						
						5 17340000
REGRESSIONS						
1	4 102024	2 104401	1-408865	2 139181	1-620519	782138
9 variables		1-713348	1 100537	1 191710	-258930	819440
8 variables 2	3 962228	2 114234	1-408863	1 687096	1-620522	782134
		1-221685	1 100537	1 191711	-258932	
3	3 985202	2 121663	1-667796	2 104657	1-102658	1 130000
		1-627040	-1-303573	622449		819443

Intervening Regressions Omitted

496	3 124417	1 856667		527778		
						-1 555555
497	3 468972	1 727055			766042	
						-1-183712
498	3 362362	1 613954				318568
						-1 836794
499	3 121644	1 916594				
		-1-918866				105936
500	3 602949	1 753291				
			178285			-2-340677
501	3 572682	1 829024				
				266332		-1-775941
502	3 672362	1 845724				
2 variables					-1 568725	-1-425946
1 variable 503	3 126000	1 773333	1 120000			
504	3 114817	1 796667		1 138333		
505	3 465121	1 713889			718889	
506	3 240944	1 781313				318568
507	3 116668	1 872759				
		334291				
508	3 602791	1 749221				
			176789			
509	3 520466	1 769914				
				202280		
510	* 3 651415	1 795792				
					-1 504621	
511	3 121419	1 905975				
						-1 836794

pendent variables. Even with the latest electronic machines, only a few of the possible formulas have usually been derived and analyzed, because of the expense of specifying hundreds of variations and then setting up individual machine inputs and programs for each.

But a new day has dawned. Now you can just send your pencil or ink data sheets (like fig. 1) or punched cards to some agency having both an IBM 704 Electronic Data Processing Machine and the Southern Forest Experiment Station's 704 Regression Program. Cost estimates will probably lie between $50 and $250, depending on the number of observations and variables, and whether you submit data sheets or punched cards. The program will completely and automatically handle up to 500 observations of Y (fig. 1 has only 30), along with observations of as many as 9 associated columns of X's (as in fig. 1). All observations should be rounded or coded so that decimals are eliminated and the number of digits does not exceed four. A supplementary program handling more than 500 observations is being developed by the Southern Forest Experiment Station and should be ready shortly, although costs will run somewhat higher.

What will you get for this relatively modest cost? See figure 2, which is only a small part of the actual output generated by the data in figure 1 (the complete program output involved 28 sheets). Briefly, the outlay buys mean Y and all mean X's; the sums of squares and cross-products of deviations from these means; the variation in Y removed by every possible formula involving one or more columns of X (up to 9 columns), and the coefficients needed in each such formula. With (m) columns of X, there will be (2^m-1) possible formulas, and they will have a total of $(m)(2^{m-1})$ coefficients plus (2^m-1) constants. For the maximum of 9 columns of X's the complete program output will furnish information concerning all of the 511 possible different formulas involving these variables, and there will be a total of 2,304 formula coefficients. There will be 1 nine-variable formula, 9 eight-variable formulas, 36 sevens, 84 sixes, 126 fives, 126 fours, 84 threes, 36 twos, and 9 ones. A table of binomial coefficients will tell you how many of each kind of formula there will be if you start with fewer than 9 columns of X's.

Figure 2 may still look like gibberish because the "floating decimal point" notation is unfamiliar. Actually, it's simple. Only the main groups of 6 or 8 digits are real quantities--the single-digit prefixes are merely instructions as to where the decimal point belongs. Where no number precedes a group of 6 or 8 digits, the decimal point is placed immediately in front of the group. Where the prefix 1 precedes the main group, the decimal point is moved one position to the right; where the prefix 2 precedes, the point is moved two positions to

the right, etc. Where -1 precedes, one zero is prefixed to the group and the decimal point is moved one place to the left (i.e., placed to the left of the zero). Where -2 precedes, two zeros are prefixed to the group and the decimal point is placed to the left of both zeros. Where no algebraic sign is printed in front of a group, plus is understood. Any minus sign is always specifically indicated. To illustrate, 2 146532 means 14.6532; 1 437986 means 4.37986; -2 -529431 means -.00529431; 764327 means .764327; -464789 means -.464789.

It would have been nice if figure 2 could have had column headings stretching horizontally from 0 through 9, after the "SUM OF SQUARES" heading. The printing machine wasn't wide enough, though, so it had to break the line after X_4 and start X_5 under the first column, or mean Y (here denoted by the 0 heading). Hence, the first numbered column-head is 0 + 5.

Now let's look at figure 2 item by item. The first double row of numbers indicates that mean Y is 10.733333, mean X_1 is 2.5000000, etc., and that each number is the mean of 30 observations. The mean of X_5 is 6.0000000, with 30 observations indicated beneath it. Note that this starts a second double row directly under mean Y. The column-head 0 + 5 indicates that the upper mean will be Y and the lower mean will be X_5. Then X_6, X_7, X_8, X_9 follow horizontally after X_5. In each case, the column-head indicates what kinds of X will be found in the upper and lower double rows (i.e., 1 + 6, 2 + 7, 3 + 8, 4 + 9).

Now we get to the part of the results subheaded "MAXIMUM MATRIX, SP and SS." All of these values are sums of squares or cross-products of deviations from some mean X's or mean Y. The particular combination is denoted by the dual column number, by the number of the double row, and by the position of the digits in the upper or lower part of the row. For example, the first entry is 1349.8667 in the upper part of row 0 under column 0 + 5. (The upper position in a row means that the first digit in the dual column-head is appropriate.) Therefore, its descriptive location is 00, which means it represents the sums of squares of deviations of Y from mean Y--often called the total sum of squares for Y. All other figures in row 0 involve cross-products of Y deviations with X deviations indicated by column-heads. In the lower part of row 3 under column 4 + 9 (descriptive location is 39), we find 2310.0000, or the sum of cross-products of deviations from mean X_3 and deviations from mean X_9. Note that the lower position in a row means that the second digit in the dual column-head is appropriate.

These figures will be helpful if you later want an inverse matrix (c-multipliers) for a few selected formulas to be calculated locally by existing programs on less expensive machines (the IBM 704 costs about $700 an hour to operate, and is too expensive for relatively routine calculations). Currently, matrix inversion (fifth- to tenth-order matrices) costs only from $30 to $35 on machines such as the IBM 650.

And now for the meat in the coconut. Each consecutive numbered double row below the subheading "REGRESSIONS" gives vital statistics about a single formula involving prediction of Y by some particular combination of the various kinds of X's. Entries in the first column to the right of the row number (the column headed "SUM OF SQUARES") are sums of squares attributable to various regressions, and any one divided by the total sum of squares (1349.8667) mentioned above gives the square of the multiple correlation coefficient (R^2) for the particular formula specified by the remainder of the double row. It is not necessary to perform this division unless you wish to make certain statistical tests or statements. However, you can readily choose the best 8-variable formula, the best 7-variable formula, etc., by merely selecting the line in each particular class which has the largest sum of squares shown in this "SUM OF SQUARES" column. As an example, figure 2 shows that asterisked regression 510 is the best of the 9 single-variable regressions; its "SUM OF SQUARES" attributable to regression (651.415) is the largest in the single-variable series 503-511.

The adjoining figures shown in double rows under columns headed 0 + 5, 1 + 6, 2 + 7, 3 + 8, and 4 + 9 are the coefficients appropriate to that particular formula, with the 0, 1, 2, 3, 4 referring to the upper position and the 5, 6, 7, 8, 9 referring to the lower position in the row. In each double row, the upper quantity shown in the 0 column is a constant. All other quantities are coefficients for X_1 through X_9. Thus, if a double row has figures appearing under 8 heads besides "SUM OF SQUARES" and "Zero" (which represents a constant), a formula is specified which depends on 8 variables to predict Y. As an example, regression # 3 is interpreted as:

$$Y = +12.1663 - 6.67796X_1 + 10.4657X_2 - 1.02658X_3 + 1.30000X_4 - 6.27040X_5 - .0303573X_6 + .622449X_7 + .819443X_9$$

For regression # 3, the sum of squared residuals (or the sum of squares of the differences between formula-predicted values for Y and the actually observed values of Y) is 1349.8667 minus 985.202, or 364.665. If desirable, the squared multiple correlation coefficient R^2 could be found as $\frac{985.202}{1349.8667} = .730$. The statistical "degrees of freedom"

appropriate to both the sum of squared residuals and to R^2 is the total number of observed sets of values (30 in this case) minus the number of coefficients used in the formula (including the constant in the "Zero" column). In the above example, degrees of freedom would be (30-9) or 21 dfs. The sum of squared residuals divided by the degrees of freedom gives the mean squared residual (or the squared standard error of estimate for the formula). In this case, the mean squared residual would be

$$\frac{364.665}{21} = 17.365 \text{ with 21 degrees of freedom.}$$

With the above explanation, it is possible to explore different sequences for taking variables into account. These techniques are well known and will not be discussed here. However, comparison of the mean squared residual with the improvement in variation progressively accounted for by the best 1-, 2-, 3-, 4-, 5-, 6-, 7-, 8-, or 9-variable formulas will often suffice to decide upon a cutoff point.

Foresters will find this program immediately helpful in converting volume tables to formulas for forest inventory punched-card operations, and in analyzing stand growth or soil-site relationships. It will also be useful wherever else multiple regression techniques are appropriate. It is far cheaper than manual or piece-meal machine calculation of the same data, and it will secure more comprehensive, convenient, and reliable information far more quickly, cheaply, and with less skilled technical knowledge than will any other existing package program the author has been able to locate or devise.

Biometrical Background

Although complex formulas fitted by least squares (i.e., multiple regressions) have gained wide favor in recent years, their use has been limited by many persons' unfamiliarity with data-processing techniques and by the expense of screening out the less efficient combinations of variables where the so-called independent variables are intercorrelated (as they usually are). The interpretation of a conventional partial correlation analysis is clouded by any intercorrelation among independent variables. In other words, estimates of the importance of a given independent variable fluctuate as the variable is grouped with different combinations of other independent variables. Hence, the path or order of fitting affects the apparent importance of any independent variable in a prediction formula. The regression analyses described in most publications deal with a single path or order of fitting, and ignore other possible sequences because of the magnitude of computational labor.

After wrestling with this problem for a number of years, the author reluctantly decided that no known existing machine program was adequate for the job, and that even such excellent equipment as the IBM Type 650 Electronic Data Processing Machine did not have the storage or the speed to allow development of the needed program with a low operational cost.

Consequently, he constructed the synthetic sample problem (designed to allow a number of cross checks) illustrated in figure 1 from D. B. DeLury's 1950 Values and Integrals of the Orthogonal Polynomials up to n = 26 (page 9), and then worked out the specifications for the program output illustrated by figure 2. Consultation with several programming agencies finally resulted in a contract with The Service Bureau Corporation, a subsidiary of IBM, to develop a program for the IBM 704 according to detailed Southern Forest Experiment Station specifications and example.

The most efficient computational approach seemed to be conversion of the maximum matrix of sums of squares and cross products of deviations to a maximum matrix of simple correlation coefficients. From this, a matrix of appropriate simple correlation coefficients (r_{ij}), always including the Y-correlations, could be constructed for each of the 511 possible regressions. Elements (C'_{ij}) of the inverse matrix for each of these correlation-coefficient matrices were computed by the modified Jordan elimination method. Then for each regression,

$$b_{oj} = -\frac{C'_{oj}}{C'_{oo}}\sqrt{\frac{SS_{oo}}{SS_{jj}}}\,,$$ with o denoting the dependent variable,

with j successively denoting each independent variable present in a given regression, and with SS denoting the sum of squared deviations from the mean. The general procedure is outlined on page 145 of C. H. Goulden's 1952 Methods of Statistical Analysis. The computation of sums of squares attributable to regression and a check on the coefficients for selected regressions are illustrated on page 137 of Goulden. Miss Carol Hadek of The Service Bureau Corporation broke down the specified mathematical processes into a staggering series of single-step machine orders permanently recorded on magnetic tape. In obedience to this tape, the IBM 704 will operate as ordered on whatever quantities may be fed into it. The Southern Forest Experiment Station has borne the expense of developing this program--a matter of several thousands of dollars. Now, however, anyone can have The Service Bureau Corporation process raw data (according to the Southern Forest Experiment Station's program) for merely the normal costs of operation ($50 to $250).

The program should be widely useful in non-forestry fields. Furthermore, similar programs can be developed in the same format to handle many more than 9 independent variables. It is hoped that other agencies will find it profitable to finance such program extensions, but the existing program will serve in the most commonly encountered forestry situations. The program outlined above, with input at the limiting maxima, is solved by the IBM 704 in about 11 minutes of actual operation. With such potentiality, there seems little need to overheat desk calculators for months on end, or to be satisfied with a single sequence of fit, as has been almost universal practice in the past.